Must NLP be Extractive?

Steven Bird
Northern Institute
Charles Darwin University
Darwin, Australia

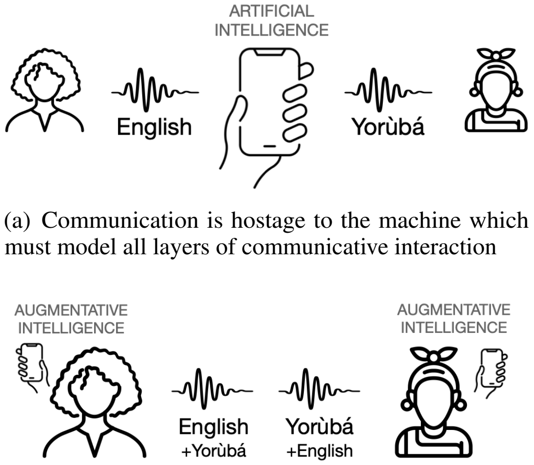
Abstract

How do we roll out language technologies across a world with 7,000 languages? In one story, we scale the successes of NLP further into 'low-resource' languages, doing ever more with less. However, this approach does not recognise the fact that – beyond the 500 institutional languages – the remaining languages are oral vernaculars. These speech communities interact with the outside world using a 'contact language'. I argue that contact languages are the appropriate target for technologies like speech recognition and machine translation, and that the 6,500 oral vernaculars should be approached differently. I share stories from an Indigenous community where local people reshaped an extractive agenda to align with their relational agenda. I describe the emerging paradigm of Relational NLP and explain how it opens the way to non-extractive methods and to solutions that enhance human agency.

1 Introduction

For over half a century this community has been developing methods for so-called 'natural' language processing (NLP). By *natural* this community does not mean the kinds of spoken interaction most people would regard as natural. We mean documents containing a textual trace of human language, as distinct from the default kind of language to be processed by computer, which is apparently programming languages. I believe that generative AI and large language models misconstrue the nature of language, and I argue that it is time for the NLP community to take 'natural language' seriously.

Meta's project "No Language Left Behind" promises to enable people to make "more *meaningful* connections in their preferred or native languages, [bringing] people together on a global scale" (Meta, 2023). Google's Universal Speech Model will "*understand* the world's 1,000 most-spoken languages" (Roth, 2023). The chatbots are going massively multilingual.



(a) Communication is hostage to the machine which must model all layers of communicative interaction

(b) Communication is amplified by the machine which provides an imperfect but helpful assistant

Figure 1: LT4All Design Patterns: machine *vs* human learning; simulating *vs* supporting humans; diminishing *vs* enhancing agency; monolingualism *vs* language mixing, translanguaging, and receptive multilingualism

So it was that I listened while an African scholar described the prospects for his friend in Switzerland to learn ancestral food practices from her grandmother in Nigeria. A translation app would solve the language barrier, he mused. I sketched the scenario (Fig. 1(a)). Yes, that's it, he said. I asked if this system would need to be trained on familial conversations with an interpreter in the middle, to be replaced by his app. Which of the 20+ dialects of Yorùbá would he pick? It would need to handle words for ingredients and implements that have no translation. And how would this system interpret the kinds of utterance that are common between family members, whose meaning depends on shared knowledge that the system has not been exposed to? We sat in silence. Yes, it's a problem, he said, and even if it was possible, it would take too long. I asked if the woman already knew some Yorùbá and if she adds it to her English. Yes, she already does that, he said, and she wants to learn more. I drew another diagram (Fig. 1(b)).

This is an essay about designing technologies for so-called 'unwritten' languages, responding to the widely-held belief that *Language Technology for All* is to be accomplished by ingesting ever more languages into massively multilingual models.

I examine the epistemics of language work that is assumed by these technology-driven approaches: *First Wave NLP* with symbolic methods and language-as-code, and *Second Wave NLP* with subsymbolic methods and language-as-data. I argue that, by treating all languages as bounded and standardised (Milroy, 2001; Krämer et al., 2022), these approaches misconstrue 'natural' language and limit the possibilities for technologies. What would a *Third Wave NLP* look like, one that centres natural language in the fullest sense of 'natural'?

I return to the fork in the road at the beginning of AI, between Artificial and Augmentative Intelligence (McCarthy, 1965; Engelbart, 1963). Is our agenda to replicate human intelligence inside a machine, or to expand human intelligence by using machines as tools? The former requires extraction of behavioural data and takes over human agency, whereas the latter carefully enhances human agency (cf. Fig. 1). What would it be like to take the other fork, and seek a path towards a non-extractive NLP committed to augmentative solutions?

By sharing stories from my time in an Indigenous community, I reveal the ingenuity of local people in repurposing extractive efforts and reshaping deficit thinking. These stories demonstrate the resilience of a minoritised speech community in pivoting from deficit to strength. They show people enacting their sovereignty and agency in shaping their lives, landscapes, and languages, in the face of the Eurocentric impulse to problematise the ways that minoritised groups fall short of western norms, beginning with the label of 'unwritten language.' This Indigenous habit of preserving agency is instructive for anyone who is ceding their agency to technology. What would an agency-enhancing NLP look like?

A promising answer to these questions that is emerging in recent work, I believe, could be called *Relational NLP*. Relationality was always implied by the communicative intent that underlies language use (Fulton, 1942). However, it cuts deeper to the way language has been passed down, how it is embedded in the land, woven into kinship, and used to accomplish more-than-physical work (Shankar and Cavanaugh, 2017; Hinton, 2022).

2 Preliminaries

Beyond the 500 institutional languages, the world's linguistic diversity consists of about 6,500 oral languages (Bird and Yibarbuk, 2024). If we are to develop technologies here, we should grasp something of the nature of linguistic diversity (Sec. 2.1), cultural diversity (Sec. 2.2); and the history of our efforts in NLP to address them (Sec. 2.3).

2.1 Language and linguistic diversity

Language is fundamentally social. We converse with intent. We hesitate and self-correct. We detect misunderstandings, interrupt, clarify. We speak our dialect using wordforms we never see written. We add new senses to old words and invent new expressions. We sign and gesture. We use intonation and facial expression to show that our words are sincere, or ironic. We understand even when we don't recognise words (Goffman, 1955; Ong, 1982; Tedlock, 1983; Sacks, 1986).

Oral languages have purely local functions; they cannot be transplanted as they are embedded in the land. People do not look up information using their oral language, they ask an elder. Parents do not transmit oral language to children using books, they tell stories. Locals do not conduct business with the outside world in the vernacular, they use the vehicular, i.e., the 'contact language', or 'trade language', or 'language of wider communication' (Basso, 1996; Fishman, 2001; Bidwell et al., 2008; Woodward and McTaggart, 2019).

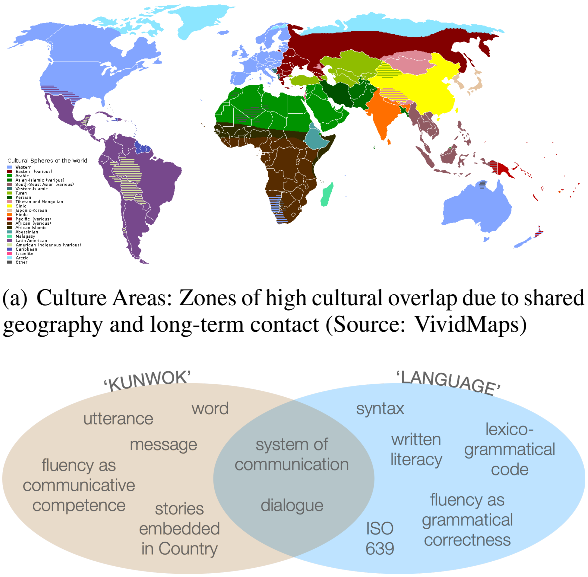
Contact languages are the legacy of imperialism, mass media, and formal education. As such, most contact languages are varieties of institutional languages, endowed with standardised orthographies, literacy, and formal education, serving the agenda of wider economic and civic participation. These languages are the obvious target for 'low-resource' language processing (Bird, 2022).

Local speech varieties are characterised by variability and mixing (Fishman, 2001; Dobrin et al., 2009; Leonard, 2017; Grosjean, 2021). Some diglossic communities practice multilingual conversation, exploiting the fact that people can recognise more than they can produce, a mode known as multilingual receptive comprehension (Asher, 1969; Davies, 1976; Meakins, 2013; Singer, 2018; Vaughan, 2021). People leverage their multilingualism in the creative practice of translanguaging (Cenoz and Gorter, 2017; Mazzaferro, 2018; Seals and Olsen-Reeder, 2020).

2.2 Culture areas and concept spaces

Culture areas are geographical regions where we find substantial cultural similarities in the midst of linguistic diversity, due to shared geography and long-term contact (Fig. 2(a); Voegelin and Voegelin 1964; Newman 1971). Traditional practices, ceremonies, and material culture are often shared across a culture area, with the result that many concepts are only lexicalised within that area (Babaii et al., 2020). "The worlds in which different societies live are distinct worlds, not merely the same world with different labels attached" (Sapir 1929, p209, cited in Hinton 2022, p60). When we transit to a new culture, many concepts are difficult to map (Wierzbicka, 1992; Evans and Sasse, 2007; Liu et al., 2021; Hershcovich et al., 2022a). The limited overlap of lexicon, semantics, and genre between oral societies and the western monoculture presents a stumbling block for machine translation.

This issue is pervasive. For example, consider the concept of *language* itself, locally considered and collectively enacted as a social practice *vs* the western notion of language as "an object isolated from interaction" (Fig. 2(b); Hermes et al. 2022, p63). In the intercultural space, translation requires exegesis (Woodbury, 2007; Lowell et al., 2021), exceeding what we can learn from parallel texts which "only address standardized, universal stories, and fail to explore what is culture-specific" (Evans and Sasse, 2007, p71).


Cultural Spheres of the World
'KUNWOK'
'LANGUAGE'
utterance
word
message
fluency as communicative competence
stories embedded in Country
system of communication
dialogue
syntax
written literacy
lexico-grammatical code
fluency as grammatical correctness
ISO 639

(a) Culture Areas: Zones of high cultural overlap due to shared geography and long-term contact (Source: VividMaps)

(b) Semantic Spaces: putative translational equivalents like 'kunwok' *language* have culturally bounded meanings

Figure 2: Culture areas as the basis for distinct semantic spaces, an alternative to a universal concept space

2.3 Three waves of NLP

First Wave NLP (1960s–1990s) consisted of rule-based approaches, within an epistemology of *language as bounded lexico-grammatical code*. Second Wave NLP (1990s–) has been characterised by statistical approaches, within an epistemology of *language as sequence data*.

In the first wave, we have responded to the challenge of linguistic diversity with linguistic software and grammar engineering (e.g. Lawler and Aristar Dry, 1998; Nirenburg, 2009). In the second wave, we have adapted machine learning methods to progressively smaller datasets (e.g. Besacier et al. 2006, 2014; Adda et al. 2016; see also Figs. 7, 8).

Despite their manifold successes, the limitations of both approaches are revealed in the way that manipulating *forms* never finally accesses *meanings*. The chatbots of the 2020s, as in the 1960s, miss out on the world (Weizenbaum, 1966; McDermott, 1976; Strauß, 2018; Natale, 2019; Bender and Hanna, 2023). Their popularity owes much to the Eliza Effect, a linguistic correlate of pareidolia, the human habit of seeing faces in clouds.

Second Wave NLP has become unsustainable (Hershcovich et al., 2022b; Morreale et al., 2023). Scraping data has violated data sovereignty (Walter and Suina, 2019; Mahelona et al., 2023). There is no onward trajectory to language understanding (Bender and Koller, 2020; Ghassemi et al., 2023; Church, 2024; Messeri and Crockett, 2024). Has Second Wave NLP run its course?

How might we get started with an NLP that embraced language as a situated and embodied social practice? We could move on from the linear, Shannon-Weaver model of communication to one which allows for the co-construction of meaning, and which sees communication and relationships as mutually constituted (Littlejohn and Foss, 2009, p177). We could take seriously the purpose of language for sustaining relationships (Eades 2013, p62; Hermes et al. 2022, p62). We could respect other relationships, such as the speech community's ownership of language (Martinez, 2000; Ting, 2023), and the Country's embedding of language (Basso, 1996; Steffensen, 2019; Hinton, 2022). We could build ethical practices on relationality (Taylor et al., 2019; Birhane, 2021; Ògúnrẹ̀mí et al., 2023; Schwartz, 2022; Carpenter et al., 2024; Bird and Yibarbuk, 2024; Cooper et al., 2024; Markl et al., 2024). But to make this concrete, we begin with stories from a place.

3 Stories from Arnhem Land

The region of Arnhem Land in northern Australia boasts several Aboriginal towns, established during the first half of the 20th century in an era when government and missions conspired to concentrate people into settlements. In a policy reversal, the 1970's homelands movement encouraged people to return to their Ancestral homelands. Today, locals move between towns and homelands during the dry season, while staying in situ during the wet season when the rivers come up, blocking the roads.

One such settlement is Kabulwarnamyo, with about 50 people situated on the 'Stone Country' of the Arnhem Plateau at 12.765°S, 133.845°E. This land belongs to the Mok clan and embeds the 'mankung djang' *honey dreaming* story. The physical setting is savannah woodland punctured with sandstone outcrops and ravines. For five months of the year, the only access to the community is via light plane. This community is the central point for the land management work of the Warddeken Indigenous ranger program (Yibarbuk et al., 2001; Garde et al., 2009; Russell-Smith et al., 2009).

One of the Warddeken rangers is Kamarrang Stuart Guymala (Fig. 3), a member of the Bordoh clan and one of the few remaining speakers of the Kundedjnjenghmi dialect of Kunwinjku. He is well respected as a fire fighter and buffalo hunter, and is a long-time member of Nabarlek, an Aboriginal band that has toured internationally.

I am descended from German and English settlers, with training in computer science and linguistics and experience of working with minoritised language groups in West Africa, Melanesia, and Amazonia. For thirty years I focussed on what technology could do for 'endangered' languages: improving orthography (Bird, 1999b), curating lexical data (Bird and Tadadjeu, 1997), 'helping' linguists (Bird, 1999a; Bird et al., 2009), preserving languages (Bird and Simons, 2003), capturing audio (Bird et al., 2014), and deployment on mobile devices (Bird, 2018). In 2015, I tried to bring this work to Aboriginal Australia, and began working in Arnhem Land where I was 'adopted' by a local family (Bird, 2016). In 2017, I came to Kabulwarnamyo to support language work in a school and ranger program, and through the patience of local people, gradually relinquished many colonial assumptions and began to take local agency and self-determination seriously (Bird, 2020).

Figure 3: The author, with principal language teacher Kamarrang (holding notebooks at Kamarrang's request)

3.1 Language work

Kunwinjku is the main language of West Arnhem, spoken by about 2,000 people. It is a polysynthetic language with complex verb morphology and noun incorporation (Evans, 2003; Lane and Bird, 2019). It has a rich vocabulary to articulate the kinship system (Garde, 2013). Kunwinjku is undergoing language shift (Marley, 2020). "Community members and elders are concerned that younger generations are not attaining a comprehensive vocabulary in Kunwinjku, which can only be reached through learning and practising language on country, as such a large percentage of the Kunwinjku vernacular is related to the natural environment" (Warddeken, 2021, p66).

To support this agenda, I began to make language resources with the help of Kamarrang. One day, I tried to extricate him from his ranger duties, with:

(1) ji-rɑ-j kɑne-ɟɑrk-durkmiri ɡun-wɔk
2-come-IMP 12-together-work IV-talk
Come and do language work with me

The rangers burst out laughing. Our 'work' was 'bimbun' *drawing* in notebooks and 'bɔŋun' *drinking* tea, whereas 'durkmiri' is ceremonial labour which honours ancestors and sits in a web of kinship obligations. Its etymology is 'durk-mi-ri' *pull-COM-stand*, the work of clearing grass and plants from ceremony ground. Its range extends to physical labour, but not desk work. 'Working' with me was viewed as time off.

Kamarrang's brother jumped up, looked at his wrist where we imagined a watch, rushed to a spot 10 metres away, sat for a moment, looked at his wrist again, and jumped up and ran to another place. Everyone laughed at white man's busy work being play-acted (cf. McRae-Williams, 2008, p188).

3.2 Guided tours for teaching from Country

The first time I recorded an audio tour, it happened on the spur of the moment, and I followed my adoptive Aboriginal sister around the community while she explained various locations and explained my language learning activities to others. This recording captured incidental participants so we did not use it. However, an idea was taking shape.

I noticed visitors being shown around the community, and I realised that guided tours were an established practice. Locals wanted to ensure that newcomers were safe and did not accidentally disturb a sacred site. I asked Kamarrang ‘kɑn-bolk-bukkɑ-∅’ 2/1-country-teach-IMP *show me around*. He objected, ‘yibɛŋgɑn gɒrɒkɒ *you know already*. I said in English, *yes, but let’s pretend*. He laughed, ‘mɑʔ ŋɑrrɒʔrɒkme’ *ok let’s try*. So I turned on the audio recorder and we set off.

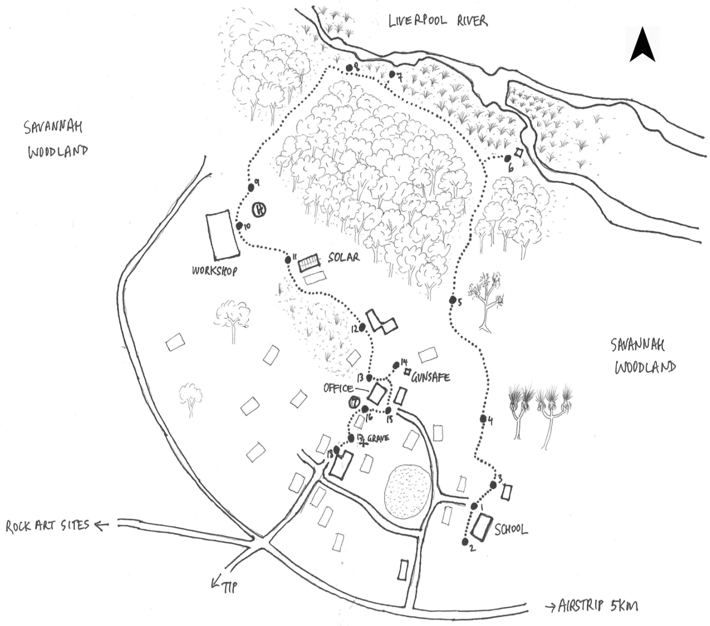
In each place, Kamarrang explained what we could see, along with its purpose and history (Fig. 4). This was Kunwinjku audio with Country as the only interpretation. No transcription or translation was needed. We talked about what we had just done. I said, I can do this with other people. . . make several recordings. He said, ‘juw’ *yes*, ‘ɟɑ bɒlkgime ŋɑrjɑwɔjʔrɒʔrɒkme kɒre gungukbɛlɛ’ *but now we’ll try again in English*. We set off once more, me as guide this time, and repeated the circuit while he pointed at things saying ‘ɲɑle nɑʔni’ *what’s that?* and I recorded plain English responses on Kamarrang’s phone.

Instructions. For each participant, I explained that we would walk around the community, and they would teach me about the place. At the start of the recording, the participant was to introduce themselves and state their clan and country. I led the way, following the same path and stopping in the same 18 locations. In places which contained plant species, I pointed and asked ‘ɲɑle nɑʔni’ *what’s that*, and people might name it and state any uses they were aware of, relevant seasons, and so on. When a word seemed significant I might parrot the word, prompting the guide to say more about it.

Recordings. Eight people were recorded as we traced the same path (Fig. 4), in 16 bit 16kHz mono using a wireless lapel microphone and a professional field recorder. Participant information is shown in Figure 5.

Spkr	Sex	Age	Duration
SG	M	50-59	10m26s
GN	M	30-39	12m20s
MM	M	30-39	07m44s
GN	M	30-39	06m10s
TG	M	50-59	09m20s
DY	M	60-69	23m30s
DM	M	20-29	15m48s
RN	F	40-49	15m35s
Total			1h40m53s

Figure 5: Participants in guided tour recordings



1. open-air school
2. school garden
3. teacher’s house
4. pandanus grove
5. woodland
6. pump house
7. spring
8. dreaming site
9. helipad
10. workshop
11. solar panels
12. visitor camp
13. ranger office
14. gunsafe
15. community hall
16. satellite dish
17. founder’s grave
18. main homestead

Figure 4: Tour of Kabulwarnamyo visiting 18 locations where locals shared their knowledge in Kunwinjku

Careful respeaking. This involves listening to an audio source phrase by phrase and repeating what was said. The original, spontaneous utterance is reproduced in careful speech in a near-field recording, as if dictating to a future transcriber (Woodbury, 2003; Abney and Bird, 2010; Sperber et al., 2013). We adapted respeaking to serve our purposes: I repeated what I heard, and Kamarrang corrected mistakes and coached me on any pieces I had missed. This was essential for morphemes that had been phonologically reduced:

(2) a. [ɡɑwdiri]
/ɡɑbiri-jɑw-diri/
3a-DIM-playing
the children are playing

b. [pɑmp ɡɑburkmɑŋ]
/pɑmp ɡɑ-bɔ-durkmɑŋ/
pump 3-liquid-suck
the pump is sucking up water

c. [ɡɑbindiɑɡɒɟuʔɡɛ]
/ɡɑbindi-jɑw-ɡɒɟ-ɟuʔɡɛ/
3a/3pl-DIM-head-splash
they're splashing water on their heads

Through this process I identified roots and affixes, and checked my interpretations. I would repeat an entire phrase, receiving further correction until I could reproduce it without error. The recording of the respeaking was for my personal use, not for other learners. To document the corpus, I transcribed everything using the Kunwinjku orthography (Fig. 6).

SG	nahne **yakngarra**no kabirribidjmang kabirrimarnbun yiman kayime **yakngarra karlba** yiman kayime kangulme **yakngarra**no ngad karriyime colour wanjh ngad karriyime **yakngarra karlba yakngarra** ngulmeng
GN	mahne man**yakngarra** kahdi, man**yakngarra** menekke daluk kabirri**ngobarn**mang kabirrimarnbun kabirrikinje bu kabirrimarnbun nawu basket ngong yiman kayime kure daluk bedberre kabirridurrkmirri
MM	mahne kun**ngobarn** kobahkohbanj kabirri**ngobarn**mang kabirrimarnbun basket. kabirri**ngobarn**yirrme wanjh kabirrinan larrakurrme wanjh kabirrikinje wanjh colour kabirrikurrme wanjh kabirri**ngobarn**njamedme basket
TG	mahne ngarringeybun man**yakngarra** man**yakngarra** kobahkohbanj kabirrimang manekke bikno karrdum menekke kahdi kadjabdi kubuldjan kabarrimang kobahkohbanj kabarriyirrme kabarrimarnbun...

Figure 6: Sample of orthographic transcriptions of recordings made in a pandanus grove (high frequency content words in boldface)

Evaluation. Every participant expressed pleasure at this activity, and that a westerner was showing respect for their knowledge, and learning about the Country. No-one was bothered that I repeated the same tour with several people. I had none of my previous issues of people absenting themselves. On the contrary, everyone was eager to participate. Local participation or non-participation in my activities was a way that I was "being participated" (Winschiers-Theophilus et al., 2010).

This task represented a turning point in my experience of living and working in the community, as it was the first time I was able to experience learner-directed speech which did not pressure me to respond in turn-taking dialogue. I was able to learn language without reliance on transcription and translation. The speech was always directed at my level of proficiency, and based on what people wanted me to learn. People mixed Kunwinjku and English, adding redundant information.

Many participants appreciated that this activity did not involve writing. The recent introduction of an orthography standard for Kunwinjku had undermined confidence: people now knew there was a correct way and that they did not know it, and so setting pen to paper risked making mistakes and experiencing shame (cf. Rehg, 2004; Hinton, 2014).

This corpus is not stand-alone, because the content is only meaningful when it is accessed at the right locations. It prompts newcomers to interact with locals in order to interpret content at their level of understanding. This aligns with how Kunwinjku learners are advised in the primer: "The authors of this book are not authorities on all matters Kunwinjku, and this book is not a bible. The best people to help you learn both culture and language are the real experts, the Kunwinjku people themselves. Our aim is to have you rely on *them* for accurate information" (Etherington and Etherington, 1998, p*i*, emphasis in original).

This corpus suggests the kind of spontaneous content that could be leveraged in augmentative intelligence scenarios (e.g. Fig. 1(b)). A mobile app could detect the speaker's location, track their proficiency by analysing recent speech productions, and suggest words and phrases that might be relevant.

This approach combines the existing ideas of learning language in the context of guided tours (Clark, 2013; Clark and Torretta, 2018), and of recording walks for language documentation and revitalisation (Cialone, 2019; Hermes et al., 2022).

4 Discussion

In order to learn what a non-extractive NLP might look like, I began from a particular Country and with the local countrymen and countrywomen who have co-existed with this Country over countless generations (Bidwell et al., 2008; Woodward and McTaggart, 2019). The details were idiosyncratic but inferences could still be made, not through induction but abduction.

Through narratives, I centred "participants' life experiences, social relationships, and observable artifacts surrounding them" (Sultana et al., 2022). In this way, "we encounter[ed] the world as a place in which we act – the practical tasks in which we are engaged, and how they are accommodated into the world – that makes the world meaningful for us" (Dourish, 2004, p108).

I could have shared other stories from this place. I chose this one because it includes something recognisable: a corpus. However, the design is relational. The corpus is inseparable from the place where it arose. The content is based on what locals want newcomers to know. It is incomplete by design, to prompt interpersonal engagement.

The corpus is an opportunity for many kinds of processing. We could collect time-aligned still images and GPS data. We could play back all commentaries relevant to a location in the context of revisiting that location. We could experience a virtual tour using the images. We could test someone's ability to produce topical words for a location. We could aggregate the tours made by different newcomers and curate the content by difficulty for future newcomers. In all such enrichments, the corpus remains non-self-contained, and it cannot meaningfully be made portable and used elsewhere. These are all possibilities for further investigation.

This work took place in an intercultural setting which exposes the principal shortcoming of the first two waves of NLP: "the assumption that there is a cross-lingual, cross-culturally common semantics to preserve fails when the common grounding does not match between cultures. Two relevant aspects here are the set of relevant *concepts*, closely identified with problems of lexicalisation, and *common sense*, i.e., the relevant propositional knowledge used in reasoning and entailment" (Hershcovich et al., 2022a, p6999, emphasis in original).

One way to address this grounding problem is to leave meaning with the embodied participants and the Country (Fig. 2(b)). We stop trying to capture every facet of context and interaction in the machine and rely on the situated embodiment of human agents (Fig. 1). We stop making "assumptions about information, communication, or technology before thinking about what development means from a community's perspective . . . imposing [global solutions] upon these cultures could threaten rather than support diversity" (Srinivasan, 2017, p214).

Those raised inside Second Wave NLP will not recognise this work as NLP. There are none of the hallmarks of empiricism, like reproducibility, or scalability, or generalisability to other places. What is more, I have described a closed corpus at a meeting with the special theme of open data.

Is it so audacious to propose a successor to empiricist NLP? Beyond the hyperbole and pareidolia, how good are large language models? Are the extractive and environmental harms really so desirable? Is translation through the mapping and re-arrangement of strings really so 'deep'? Is responding to the world's linguistic diversity with more of the same really so imaginative?

More than this even, the agenda of *Language Technology for All* rests on dubious assumptions:

1. that language technologies must be capable of simulating human communication;
2. that the Eurocentric practice of delimiting languages should be applied globally;
3. that all languages should be standardised;
4. that all languages have a standard orthography or would benefit from one;
5. that vernacular language literacy is universal, or universally desirable;
6. that all people are monolingual and use a single language for all communicative functions;
7. that all people use pure language, not routinely mixing vernaculars, or mixing the vernacular with the vehicular;
8. that human communication is adequately represented by the noisy-channel model;
9. that language technology scalability requires one-size-fits-all solutions; and
10. that sufficient manipulation of linguistic forms will ultimately arrive at meaning.

Big Tech promises to connect everyone, but theirs is a "world of shallow diversity" (Srinivasan, 2017, p215). Big Tech claims massively multilingual text translation will improve lives and save languages and cultures (NLLB Team, 2024, p5). Big Tech delivers text to people who do not read or write their oral language. And so, with small, non-standardised, genre-mismatched data, translation will be poor and technology disparities will be amplified (Toyama, 2015; Galla, 2016). Big Tech will tell minoritised communities to provide more data if they want better language technologies.

In view of the risks brought about by projecting the template of institutional languages onto the world's oral societies, I believe that we need to establish *Critical NLP*, with the goal of exposing and challenging the power structures enacted by Big Tech (cf. Srinivasan, 2017, pp208f).

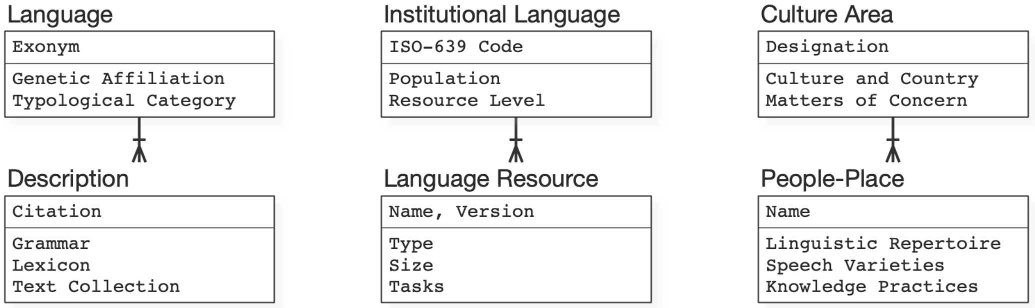
There is plenty of work for Second Wave NLP in the space of institutional languages. Recall that contact languages are institutional languages by virtue of the history of colonial contact and mass media. If the goal is *Language Technology for All People*, we need look no further than the 500 institutional languages in all their varieties, with regional pronunciation, lexicon, and grammar, often poorly supported even for dialects of English (Blodgett et al., 2020; Markl and Lai, 2021).

New prospects are suggested by Third Wave NLP: recasting language technologies as tools in service to human agency, and building the linguistic capacity of humans (Fig. 1(b); Lothian et al. 2019; Steffensen 2019; Brinklow 2021; Meighan 2021); offering translation within culture areas where lexico-grammatical methods may suffice thanks to the shared lexicalisation of concepts (cf. "zones of translatability" Bird 2022, Fig. 2); moving beyond lexico-grammatical translation to *thick translation* (Evans and Sasse, 2007; Woodbury, 2007); seeing dialects as opportunities (Nigmatulina et al., 2020; Markl et al., 2023); learning from HCI (Harrison et al., 2011; Hardy et al., 2019; Taylor et al., 2019); rethinking language technologies as sociotechnical systems (Bow, 2019; Santy et al., 2021); and developing methods that unravel the attachment to "structural properties at the expense of social practices" (Leonard 2017, p18; Barcham 2023).

I conclude by summarising the three waves of NLP along seven dimensions (Fig. 7).

First Wave NLP: **Symbolic Language Processing** Centering formal language theory	Second Wave NLP: **Subsymbolic Language Processing** Centering machine learning	Third Wave NLP: **Relational Language Processing** Centering human agency
Epistemology: • language as lexico-grammatical code • a set of well-formed sentences • modules from phonetics to pragmatics • universal grammar • realism	Epistemology: • language as sequence data • standard orthographies • primary textuality • noisy-channel model • empiricism	Epistemology: • language as embodied social practice • mixed and evolving language varieties • primary orality • co-construction of meaning • constructivism
Axiology: • quality, balance, precision, coverage • descriptive and theoretical adequacy	Axiology: • quantity, generality, efficiency • scalable technological solutions	Axiology: • agency, self-determination • sustainability, degrowth
Teleology: • linguistic description • linguistic theory and typology • natural language interfaces	Teleology: • accessing the world's information • unlocking knowledge • saving languages	Teleology: • cultural survival • knowledge transmission • healthy Country
Ideologies: • boundedness, binary grammaticality • adult monolinguals, purism	Ideologies: • open, reusable, portable data • standard language	Ideologies: • local ownership of language • embedding of language in the Country
Problematisations: • the world's languages in crisis • missing typological datapoints • language shift, mixing • loss to science	Problematisations: • low-resource languages, data gaps • language barriers • unwritten languages • inconsistent spelling	Problematisations: • loss of traditional identity • loss of ecological knowledge • loss of wellbeing • orphaned Country
Methods: • controlled elicitation • induction of theory	Methods: • data capture • induction of models	Methods: • story work, right people / place / time • abduction to underlying causes
Resources: • electronic text collections • machine-readable lexicons • formal grammars	Resources: • large corpora • large language models • language technologies	Resources: • elders • Country • linguistic repertoire

Figure 7: The epistemics of language work: Three waves of Natural Language Processing



(a) Language as code (b) Language as data (c) Language as social practice

Figure 8: Epistemics of language work (Fig. 7) reduced to ontologies and shoehorned into E-R diagrams

5 Conclusion

The agenda of *Language Technology for All* risks perpetuating epistemic harm by centering the expectation that language is primarily textual, standardised, institutional. This misses the social function of language and renders speech communities invisible (Leonard, 2017; Hermes et al., 2022). There is no path from textual sequences to *meaning*, when "meaning is irreducibly connected to the viewpoints, interactions, histories, and local resources available to those making sense of the interface and therefore to some extent beyond the reach of formalization" (Harrison et al., 2011, p388). "We need to understand the sociocultural contexts of speakers... Language is impossible to separate from context – it is continually both reflecting and creating aspects of context" (Eades, 2013, p57).

I stumbled upon a non-extractive NLP through an extended engagement with an Indigenous community that still inhabits their ancestral Country, and who resisted work that sought to capture, commodify, and carry off their language. I came to see my attachment to technological solutions as deeply problematic. I relinquished this agenda and began learning to speak the local vernacular, participating in the local lifeworld, and grasping what 'language' is and does locally. I started asking: what's happening with the children, the climate, the Country (cf. Aquino et al., 2024; Bird et al., 2024).

The theme of this ACL meeting on *open data* arises in a period where researchers seek to extend NLP to the next thousand languages (Mariani, 2020; Bapna et al., 2022; Javed et al., 2022). How can we scale NLP without open data?

This is second wave thinking. We need to approach oral vernaculars differently. We need to theorise the new paradigm of *Relational NLP* with an authentic and grounded notion of language. The notation in Figure 8 gives an inkling of the kind of shift that I have in mind, though it over-simplifies things. In each local place where we seek to deliver language technologies, we need to ask: what *is* language here, what does language *do* here? "We should recognize the importance of designing technologies with the cultures and communities they are supposed to serve [so] they can support not only the local ontologies and voices of these peoples but empower performances and practices that bind and sustain community" (Srinivasan, 2017, p210).

The defence of Third Wave NLP rests on the need to take *natural* language seriously. We reexamine the nature of language, linguistic diversity, and human communication. The stories I shared represent one possible way of many. Indigenous spaces are full of them, thanks to the habit of learning from the Country and guarding human agency. We need but look.

Acknowledgements

I am grateful to the Bininj people for accepting me into West Arnhem society and for 'participating' me in their life ways. I am profoundly indebted to Kamarrang Stuart Guymala for his friendship and generosity. I gratefully acknowledge comments from audiences where I presented earlier versions of this work, at SOAS, U Edinburgh, LORIA, Paris Cité, ENS, Orléans, Google, and Stanford. Many thanks to Teresa Lynn, Nay San, Helen Verran, and three anonymous reviewers for their thoughtful comments on an earlier version of this paper.

This work was supported by two federal grants: *Mobile Software for Oral Language Learning in Arnhem Land* (Indigenous Language and Arts Program, 2018-2021), and *Learning English and Aboriginal Languages for Work* (Australian Research Council, 2019-24). The research has been approved by traditional custodians, the board of Warddeken Land Management, the CDU Human Research Ethics Committee, and the Northern Land Council.

Ethical Considerations

Mitigating harm

The design of the guided tour activity considered many risks of harm, just as there is in any data collection work that seeks to quantify and capture Indigenous populations and their knowledge (Kukutai and Taylor, 2016). Of paramount importance was that participants and community leaders would understand how the data would be used in order to give their informed consent. Our only way to be sure of the uses was to keep the corpus private and only allow for its use in the community where it was recorded. Further risks were examined and mitigated, as follows:

Capturing incidental speech: Recording was conducted away from other members of the community, by navigating the perimeter and avoiding interactions when coming to the centre at the end.

Exposing lack of knowledge: At each location, I asked the guide in English or in Kunwinjku "what's this?" or "tell me about this place". I avoided questions that might have been received as testing people's knowledge, exposing gaps.

Visiting a secret or sacred place: The route was approved by Bulanj Dean Yibarbuk, a senior elder, and the same route was used for all tours.

Wasting a local person's time: Participation as a guide was optional and it was compensated. All participants were enthusiastic to serve as the guide. There was no concern that I may have already grasped some of the knowledge of the various places visited, such that it was pointless to refresh my memory or tell me more. On the contrary, people had observed others giving me tours and were interested to give their own version.

Boring locals with menial work: Only one tour was recorded per guide, except in one instance where the guide asked to redo the tour. The task was always novel for participants.

Recolonising behaviours: The focus of the activity was knowledge of the Country, not valorising or commodifying the language.

Pressuring people to participate: There was no compulsion to participate, and all participants were happy to share their knowledge in connection with our ongoing relationship.

Informed Consent

All members of the community were well acquainted with me, given my extended presence and work in the ranger program and the school. They knew their kinship relationship to me, and the terms of address to be used. They were familiar with my responsibilities in supporting language work in the school and the ranger program. Everyone knew that I was working under the supervision of Bulanj Dean Yibarbuk, a senior elder, who had been appointed in his leadership role by the founder of the community a decade earlier.

Everyone was familiar with the existing practice of giving guided tours to visitors, including government officials and diplomats, and the associated prestige. As a case in point, Kamarrang asked me to give him the tour in Australian English, so that he could do this himself, instead of the default Aboriginal English which carries lower prestige.

Before each tour, I reviewed the following points with each participant:

1. my reason for being present in the community, supporting language activities;
2. my purpose in recording guided tours in order to support my learning and that of other newcomers;
3. my interest in learning about the Country so that I and other newcomers would behave appropriately in this place, and that newcomers and locals could work together more easily;
4. that I wanted to record the tours so that I and other newcomers could repeat the tour later and hear the same teaching multiple times;
5. that I would put the recordings on mobile phones so other newcomers could listen as they did the tour, and learn about the place;
6. that the participant should not disclose secret knowledge;
7. that the work would be compensated at the standard rate for cultural work; and
8. that they could stop the activity at any time, or repeat it another day.

After the tour, I checked that each participant was happy with what we did. One man asked to redo the tour, and so we did this.

References

Steven Abney and Steven Bird. 2010. The Human Language Project: Building a universal corpus of the world's languages. In *Proceedings of the 48th Meeting of the Association for Computational Linguistics*, pages 88–97.

Gilles Adda, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélène Bonneau-Maynard, Elodie Gauthier, Pierre Godard, Fatima Hamlaoui, Dmitry Idiatov, Guy-Noël Kouarata, Lori Lamel, Emmanuel-Moselly Makasso, Annie Rialland, Sebastian Stüker, Mark Van de Velde, François Yvon, and Sabine Zerbian. 2016. Innovative technologies for under-resourced language documentation: The BULB Project. In *Workshop on Collaboration and Computing for Under Resourced Languages, International Conference on Language Resources and Evaluation*, pages 59–66. ELRA.

Angelina Aquino, Ian Mongunu Gumbula, Nicola Bidwell, and Steven Bird. 2024. What's the weather story? both-ways learning in Indigenous-led climate communication workshops in northern Australia. In *Participatory Design Conference*. ACM.

James Asher. 1969. The total physical response approach to second language learning. *Modern Language Journal*, 53:3–17.

Esmat Babaii, Mahmood Reza Atai, and Abbas Parsazadeh. 2020. A call for international recognition of culture-specific words from the Middle East. *Asian Englishes*, 22:106–110.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Nikhil Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Saldinger Axelrod, Jason Riesa, Yuan Cao, Mia Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apu Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Richard Hughes. 2022. Building machine translation systems for the next thousand languages. Technical report, Google Research.

Manuhuia Barcham. 2023. Towards a radically inclusive design: Indigenous story-telling as codesign methodology. *CoDesign*, 19:1–13.

Keith Basso. 1996. *Wisdom Sits in Places: Landscape and language among the Western Apache*. UNM Press.

Emily Bender and Alex Hanna. 2023. AI causes real harm: Let's focus on that over the end-of-humanity hype. *Scientific American*.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

Laurent Besacier, Bowen Zhou, and Yuqing Gao. 2006. Towards speech translation of non written languages. In *Spoken Language Technology Workshop*, pages 222–225. IEEE.

Nicola Bidwell, Peta-Marie Standley, Tommy George, and Vicus Steffensen. 2008. The landscape's apprentice: lessons for place-centred design from grounding documentary. In *Proceedings of the 7th Conference on Designing Interactive Systems*, pages 88–98. ACM.

Steven Bird. 1999a. Multidimensional exploration of online linguistic field data. In Pius Tamanji, Masako Hirotani, and Nancy Hall, editors, *Proceedings of the 29th Annual Meeting of the Northeast Linguistics Society*, pages 33–47. GLSA, University of Massachussetts at Amherst.

Steven Bird. 1999b. When marking tone reduces fluency: An orthography experiment in Cameroon. *Language and Speech*, 42:83–115.

Steven Bird. 2016. Computing gives us tools to preserve disappearing languages. *The Conversation*, June. https://theconversation.com/computing-gives-us-tools-to-preserve-disappearing-languages-60235.

Steven Bird. 2018. Designing mobile applications for endangered languages. In *Oxford Handbook of Endangered Languages*, pages 842–861. Oxford University Press.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, page 3504–3519.

Steven Bird. 2022. Local languages, third spaces, and other high-resource scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 7817—-7829.

Steven Bird, Angelina Aquino, and Ian Mongunu Gumbula. 2024. Envisioning NLP for intercultural climate communication. In *Natural Language Processing meets Climate Change*.

Steven Bird, Lauren Gawne, Katie Gelbart, and Isaac McAlister. 2014. Collecting bilingual audio in remote indigenous communities. In *Proceedings of the 25th International Conference on Computational Linguistics*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

Steven Bird and Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language*, 79:557–82.

Steven Bird and Maurice Tadadjeu. 1997. *Petit Dictionnaire Yémba-Français (Dschang-French Dictionary)*. Cameroon: ANACLAC.

Steven Bird and Dean Yibarbuk. 2024. Centering the speech community. In *Proceedings of the 18th Conference of the European Association for Computational Linguistics*, pages 826–839.

Abeba Birhane. 2021. Algorithmic injustice: a relational ethics approach. *Patterns*, 2:1–9.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

Catherine Bow. 2019. Diverse socio-technical aspects of a digital archive of Aboriginal languages. *Archives and Manuscripts*, 47:94–112.

Nathan Thanyehténhas Brinklow. 2021. Indigenous language technologies: Anti-colonial oases in a colonizing (digital) world. *International Journal of Indigenous Education Scholarship*, 1:239–266.

Craig John Carpenter, John Lyon, Miles Thorogood, and Jeannette C. Armstrong. 2024. Seeding alignment between language technology and indigenous methodologies: A decolonizing framework for endangered language revitalization. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages*, pages 318–324.

Jasone Cenoz and Durk Gorter. 2017. Minority languages and sustainable translanguaging: Threat or opportunity? *Journal of Multilingual and Multicultural Development*, 38:901–912.

Kenneth Church. 2024. Emerging trends: When can users trust GPT, and when should they intervene? *Natural Language Engineering*, 30:417–427.

Claudia Cialone. 2019. *Placing spatial language and cognition in context through an investigation of Bininj Kunwok navigation talk*. Ph.D. thesis, Australian National University.

Brendon Clark. 2013. Generating publics through design activity. In Wendy Gunn, Ton Otto, and Rachel Smith, editors, *Design Anthropology: Theory and Practice*. Bloomsbury.

Brendon Clark and Nicholas Torretta. 2018. *Co-creating language learning journeys: A designerly approach to supporting experiential language learning practices*. Barcelona: KONECT Project.

Ned Cooper, Courtney Heldreth, and Ben Hutchinson. 2024. "it's how you do things that matters": Attending to process to better serve indigenous communities with language technologies. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–211. Association for Computational Linguistics.

Norman F. Davies. 1976. Receptive versus productive skills in foreign language learning. *The Modern Language Journal*, 60:440–443.

Lise Dobrin, Peter Austin, and David Nathan. 2009. Dying to be counted: The commodification of endangered languages in documentary linguistics. *Language Documentation and Description*, 6:37–52.

Paul Dourish. 2004. *Where the Action Is: The Foundations of Embodied Interaction*. MIT Press.

Diana Eades. 2013. They don't speak an Aboriginal language, or do they? In Diana Eades, editor, *Aboriginal Ways of Using English*, pages 56–75. Aboriginal Studies Press.

Douglas C. Engelbart. 1963. A conceptual framework for the augmentation of man's intellect. In Paul W. Howerton and David C. Weeks, editors, *Vistas in Information Handling*, pages 1–29. Spartan Books.

Steve Etherington and Narelle Etherington. 1998. *Kunwinjku Kunwok: A Short Introduction to Kunwinjku Language and Society*, 3rd edition. Gunbalanya: Kunwinjku Language Centre.

Nicholas Evans. 2003. *Bininj Gun-wok: A Pan-Dialectal Grammar of Mayali, Kunwinjku and Kune*. Pacific Linguistics. Australian National University.

Nicholas Evans and Hans-Jürgen Sasse. 2007. Searching for meaning in the Library of Babel: field semantics and problems of digital archiving. *Language Documentation and Description*, 4:58–99.

Joshua A. Fishman. 2001. Why is it so hard to save a threatened language? In Joshua A. Fishman, editor, *Can Threatened Languages be Saved?: Reversing Language Shift, Revisited: a 21st Century Perspective*, pages 1–22. Multilingual Matters.

James Street Fulton. 1942. Our knowledge of one another. *The Philosophical Review*, 51:456–475.

Candace Kaleimamoowahinekapu Galla. 2016. Indigenous language revitalization, promotion, and education: Function of digital technology. *Computer Assisted Language Learning*, 29:1137–1151.

Murray Garde. 2013. *Culture, Interaction and Person Reference in an Australian Language: An ethnography of Bininj Gunwok communication*. John Benjamins.

Murray Garde, Bardayal Lofty Nadjamerrek, Mary Kolkkiwarra, Jimmy Kalarriya, Jack Djandjomerr, Bill Birriyabirriya, Ruby Bilindja, Mick Kubarkku, and Peter Biless. 2009. The language of fire: seasonality, resources and landscape burning on the Arnhem Land Plateau. In *Culture, Ecology and Economy of Fire Management in North Australian Savannas: Rekindling the Wurrk Tradition*. CSIRO Publishing.

Marzyeh Ghassemi, Abeba Birhane, Mushtaq Bilal, Siddharth Kankaria, Claire Malone, Ethan Mollick, and Francisco Tustumi. 2023. ChatGPT one year on: who is using it, how and why? *Nature*, 624:39–41.

Erving Goffman. 1955. On face-work: An analysis of ritual elements in social interaction. *Psychiatry*, 18:213–231.

François Grosjean. 2021. *Life as a bilingual: Knowing and using two or more languages*. Cambridge University Press.

Jean Hardy, Susan Wyche, and Tiffany Veinot. 2019. Rural HCI research: Definitions, distinctions, methods, and opportunities. *Proceedings of the ACM Conference on Human-Computer Interaction*, 3:1–33.

Steve Harrison, Phoebe Sengers, and Deborah Tatar. 2011. Making epistemological trouble: Third-paradigm HCI as successor science. *Interacting with Computers*, 23:385–392.

Mary Hermes, Mel Engman, James McKenzi, and Meixi. 2022. What documenting for reclamation looks like: Ojibwemowin forest walks. In Sarah Sandman, Shannon Bischoff, and Jens Clegg, editors, *Voices: Perspectives from the International Year of Indigenous Languages*, Language Documentation and Conservation Special Publication 27, pages 62–74. University of Hawai‘i Press.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022a. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 6997–7013.

Daniel Hershcovich, Nicolas Webersinke, Mathias Kraus, Julia Bingler, and Markus Leippold. 2022b. Towards climate awareness in NLP research. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2480–2494.

Leanne Hinton. 2014. Orthography wars. In Michael Cahill and Keren Rice, editors, *Developing Orthographies for Unwritten Languages*, pages 139–168. SIL International.

Leanne Hinton. 2022. *Flutes of Fire: An Introduction to Native California Languages*, 2nd edition. Heyday Books.

Tahir Javed, Sumanth Doddapaneni, Abhigyan Raman, Kaushal Santosh Bhogale, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M Khapra. 2022. Towards building ASR systems for the next billion users. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10813–21.

Philipp Krämer, Ulrike Vogl, and Leena Kolehmainen. 2022. What is "language making"? *International Journal of the Sociology of Language*, 274:1–27.

Tahu Kukutai and John Taylor, editors. 2016. *Indigenous data sovereignty: Toward an agenda*. ANU Press.

William Lane and Steven Bird. 2019. Towards a robust morphological analyzer for Kunwinjku. In *Proceedings of the 17th Annual Workshop of the Australasian Language Technology Association*, pages 1–9.

John M. Lawler and Helen Aristar Dry, editors. 1998. *Using Computers in Linguistics*. London: Routledge.

Wesley Y Leonard. 2017. Producing language reclamation by decolonising 'language'. *Language Documentation and Description*, 14:15–36.

Stephen W. Littlejohn and Karen A. Foss. 2009. *Encyclopedia of Communication Theory*. Sage Publications.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485. Association for Computational Linguistics.

Delaney Lothian, Gökçe Akçayır, and Carrie Demmans Epp. 2019. Accommodating indigenous people when using technology to learn their ancestral language. *Contexts*, 11:19.

Anne Lowell, Elaine Läwurrpa Maypilama, and Rosemary Gundjarranbuy. 2021. Finding a pathway and making it strong: Learning from Yolŋu about meaningful health education in a remote Indigenous Australian context. *Health Promotion Journal of Australia*, 32:166–178.

Keoni Mahelona, Gianna Leoni, Suzanne Duncan, and Miles Thompson. 2023. OpenAI's whisper is another case study in colonisation. *Papa Reo*.

Joseph J Mariani. 2020. Language technology for all: a challenge. In *UNESCO Report on Languages*.

Nina Markl, Lauren Hall-Lew, and Catherine Lai. 2024. Language technologies as if people mattered: Centering communities in language technology development. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 10085–99.

Nina Markl and Catherine Lai. 2021. Context-sensitive evaluation of automatic speech recognition: considering user experience and language variation. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 34–40. ACL.

Nina Markl, Electra Wallington, Ondrej Klejch, Thomas Reitmaier, Gavin Bailey, Jennifer Pearson, Matt Jones, Simon Robinson, and Peter Bell. 2023. Automatic transcription and (de)standardisation. In *Proceedings of the 2nd Annual Meeting of the Special Interest Group on Under-resourced Languages*, pages 93–97. ELRA/ISCA.

Alexandra Marley. 2020. *Kundangkudjikaberrk: Variation and change in Bininj Kunwok, a Gunwinyguan language of Northern Australia*. Ph.D. thesis, Australian National University.

Rebecca Blum Martinez. 2000. Languages and tribal sovereignty: Whose language is it anyway? *Theory into Practice*, 39:211–219.

Gerardo Mazzaferro. 2018. *Translanguaging as everyday practice*, volume 28 of *Multilingual Education*. Springer.

John McCarthy. 1965. Plans for the Stanford Artificial Intelligence Project. https://exhibits.stanford.edu/stanford-pubs/catalog/zb170hc7251, accessed May 2024.

Drew McDermott. 1976. Artificial intelligence meets natural stupidity. *ACM SIGART Bulletin*, 57:4–9.

Eva McRae-Williams. 2008. *Understanding 'work' in Ngukurr: A Remote Australian Aboriginal Community*. Ph.D. thesis, Charles Darwin University.

Felicity Meakins. 2013. Mixed languages. In Peter Bakker and Yaron Matras, editors, *Contact languages: A comprehensive guide*, pages 159–228. Mouton De Gruyter.

Paul J Meighan. 2021. Decolonizing the digital landscape: The role of technology in indigenous language revitalization. *AlterNative: An International Journal of Indigenous Peoples*, 17:397–405.

Lisa Messeri and M. J. Crockett. 2024. Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002):49–58.

Meta. 2023. No Language Left Behind: Driving inclusion through the power of AI translation. https://ai.meta.com/research/no-language-left-behind/, accessed May 2024.

James Milroy. 2001. Language ideologies and the consequences of standardization. *Journal of Sociolinguistics*, 5:530–555.

Fabio Morreale, Elham Bahmanteymouri, Brent Burmester, Andrew Chen, and Michelle Thorp. 2023. The unwitting labourer: Extracting humanness in AI training. *AI and Society*, pages 1–11.

Simone Natale. 2019. If software is narrative: Joseph Weizenbaum, artificial intelligence and the biographies of ELIZA. *New Media and Society*, 21:712–728.

James Newman. 1971. The culture area concept in anthropology. *Journal of Geography*, 70:8–15.

Iuliia Nigmatulina, Tannon Kew, and Tanja Samardzic. 2020. ASR for non-standardised languages with dialectal variation: the case of Swiss German. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 15–24.

Sergei Nirenburg, editor. 2009. *Language Engineering for Lesser-Studied Languages*. IOS Press.

NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*. https://doi.org/10.1038/s41586-024-07335-x.

Tolúlọpẹ́ Ògúnrẹ̀mí, Wilhelmina Onyothi Nekoto, and Saron Samuel. 2023. Decolonizing NLP for "low-resource languages": Applying Abebe Birhane's relational ethics. *Global Review of AI Community Ethics*, 1(1).

Walter Ong. 1982. *Orality and Literacy: The Technologizing of the Word*. Routledge.

Kenneth L Rehg. 2004. Linguists, literacy, and the law of unintended consequences. *Oceanic Linguistics*, 43:498–518.

Emma Roth. 2023. Google's one step closer to building its 1,000-language AI model. https://www.theverge.com/2023/3/6/23627788/google-1000-language-ai-universal-speech-model, accessed May 2024.

Jeremy Russell-Smith, Peter Whitehead, and Peter Cooke. 2009. *Culture, ecology and economy of fire management in North Australian savannas: rekindling the Wurrk tradition*. CSIRO Publishing.

Oliver Sacks. 1986. The president's speech. In Barbara Mayor and A. K. Pugh, editors, *Language, Communication and Education*, pages 23–27. Routledge.

Sebastin Santy, Kalika Bali, Monojit Choudhury, Sandipan Dandapat, Tanuja Ganu, Anurag Shukla, Jahanvi Shah, and Vivek Seshadri. 2021. Language translation as a socio-technical system: Case-studies of mixed-initiative interactions. In *Proceedings of the 4th ACM SIGCAS Conference on Computing and Sustainable Societies*, pages 156–172. ACM.

Edward Sapir. 1929. The status of linguistics as a science. *Language*, pages 207–214.

Lane Schwartz. 2022. Primum Non Nocere: Before working with Indigenous data, the ACL must confront ongoing colonialism. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 724–731.

Corinne Seals and Vincent Olsen-Reeder. 2020. Translanguaging in conjunction with language revitalization. *System*, 92:102277.

Shalini Shankar and Jillian R Cavanaugh. 2017. Toward a theory of language materiality: An introduction. In *Language and materiality: Ethnographic and theoretical explorations*, pages 1–28. Cambridge University Press.

Ruth Singer. 2018. A small speech community with many small languages: The role of receptive multilingualism in supporting linguistic diversity at Warruwi Community (Australia). *Language and Communication*, 62:102–118.

Matthias Sperber, Graham Neubig, Christian Fügen, Satoshi Nakamura, and Alex Waibel. 2013. Efficient speech transcription through respeaking. In *Proceedings of the 14th Annual Conference of the International Speech Communication Association*, pages 1087–91.

Ramesh Srinivasan. 2017. *Whose Global Village?: Rethinking how Technology Shapes Our World*. NYU Press.

Victor Steffensen. 2019. Putting the people back into the country. In Jo ann Archibald Q'um Q'um Xiiem, Jenny Bol Jun Lee-Morgan, and Jason De Santolo, editors, *Decolonizing Research: Indigenous Storywork as Methodology*, pages 224–238. Bloomsbury Publishing.

Stefan Strauß. 2018. From big data to deep learning: A leap towards strong AI or 'intelligentia obscura'? *Big Data and Cognitive Computing*, 2:16.

Sharifa Sultana, Renwen Zhang, Hajin Lim, and Maria Antoniak. 2022. Narrative datasets through the lenses of NLP and HCI. In *Proceedings of the Second Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 47–54.

Jennyfer Lawrence Taylor, Wujal Wujal Aboriginal Shire Council, Alessandro Soro, Paul Roe, and Margot Brereton. 2019. A relational approach to designing social technologies that foster use of the Kuku Yalanji language. In *Proceedings of the 31st Australian Conference on Human-Computer-Interaction*, pages 161–172.

Dennis Tedlock. 1983. *The Spoken Word and the Work of Interpretation*. University of Pennsylvania Press.

Chien Ju Ting. 2023. The discursive construction of language ownership and responsibility for Indigenous language revitalisation. *Journal of Sociolinguistics*, 28:46–64.

Kentaro Toyama. 2015. *Geek heresy: Rescuing social change from the cult of technology*. PublicAffairs.

Jill Vaughan. 2021. Enduring and contemporary code-switching practices in Northern Australia. *Languages*, 6:90.

Charles F Voegelin and Florence Marie Voegelin. 1964. Languages of the world: Native America, fascicle one. *Anthropological Linguistics*, 6(6):1–149.

Maggie Walter and Michele Suina. 2019. Indigenous data, indigenous methodologies and indigenous data sovereignty. *International Journal of Social Research Methodology*, 22:233–43.

Warddeken. 2021. *Warddeken Annual Report 2020–21*. Warddeken Land Management Limited.

Joseph Weizenbaum. 1966. ELIZA: A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9:36–45.

Anna Wierzbicka. 1992. *Semantics, culture, and cognition: Universal human concepts in culture-specific configurations*. Oxford University Press.

Heike Winschiers-Theophilus, Shilumbe Chivuno-Kuria, Gereon Koch Kapuire, Nicola Bidwell, and Edwin Blake. 2010. Being participated: a community approach. In *Proceedings of the 11th Biennial Participatory Design Conference*, pages 1–10. ACM.

Anthony C. Woodbury. 2003. Defining documentary linguistics. *Language Documentation and Description*, 1:35–51.

Anthony C. Woodbury. 2007. On thick translation in linguistic documentation. *Language Documentation and Description*, 4:120–35.

Emma Woodward and Patricia Marrfurra McTaggart. 2019. Co-developing Indigenous seasonal calendars to support 'healthy country, healthy people' outcomes. *Global Health Promotion*, 26:26–34.

D. Yibarbuk, P.J. Whitehead, J. Russell-Smith, D. Jackson, C. Godjuwa, A. Fisher, P. Cooke, D. Choquenot, and D.M.J.S. Bowman. 2001. Fire ecology and Aboriginal land management in central Arnhem Land, northern Australia: A tradition of ecosystem management. *Journal of Biogeography*, 28:325–343.